# Definition

## Project Overview

A key component to any strategic marketing, branding or business growth is market segmentation. The data set analyzed in this Capstone project is a user data collection from a mobile marketplace app for used goods. By segmenting the sellers into multiple groups, the company could provide a better customer support by training support staffs accordingly for each seller group when sellers reach out for assist throughout the process of selling their items.

## Problem Statement

The goal of the project is to segment the sellers into number of groups and investigate the characteristics and uniqueness of each group. The preliminary assessment by the data provider suggests that there are four distinct seller groups: top sellers, business sellers, casual sellers and new sellers.

This capstone project will further investigate the data using various techniques of clustering analysis and will determine the number of unique seller groups based on the given features of the data set.

## Metrics

Silhouette analysis will be used to test the number of clusters after KMeans clustering analysis was conducted on the dataset. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. (scikit learn webpage)

# Analysis

## Data Exploration

The dataset analyzed in this project contains sellers’ activities which are shown as columns (features) in the form of csv file. The descriptions of features are shown in Table XX.

|  |  |
| --- | --- |
| **Column name** | **Description** |
| id | user ID |
| install\_date | the user install date |
| time\_on\_site | days since the user install date |
| positive\_rating | number of positive ratings the user received as a seller |
| neutral\_rating | number of neutral ratings the user received as a seller |
| negative\_rating | number of negative ratings the user received as a seller |
| listing | number of items the user has listed for sale |
| listing\_gmv | total dollar amount of the listed items from the user |
| sale | number of sales the user made |
| buyers | number of unique buyers of the user’s items |
| gmv | total dollar amount of the user’s sold item |

**Explain about data value of zero**

Though the length of the dataset contains over 1.2 million historic seller records, majority of them implies new users who did not show any sale history using the app (‘sale’ column in the data set is zero and therefore other columns as well) at the time of this dataset. Only 8.07% of the dataset represents the sellers with at least one item listed for sale (97,423 out of total 1,207,774 sellers). These users are referred to as ‘active users’ in the following sections throughout the analysis.

**Distribution of active sellers**

Scatter plots and histograms are shown in Figure XX, which shows very wide ranges of the users across plotted features. The plots also represent the skewedness of data toward lower quantities for the features such as ‘listing,’ ‘sale’ and ‘gmv’ (total revenue). We can deduce based on these plots that majority of the sellers in the dataset represents new or relatively casual sellers. Throughout this project, the focus will be given to segregate users beyond the new and casual levels, the definitions of which will also be determined by further analysis.

**Histogram for price or item? Or some way to visualize the distributions**

Explain abnormalities in data

Show variance and skewedness of each feature

* If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?
* If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?
* ~~If a dataset is not present for this problem, has discussion been made about the input space or input data for your problem?~~
* Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)

## Exploratory Visualization – this needs to be revisited later

* Have you visualized a relevant characteristic or features about the dataset or input data?
* Is the visualization thoroughly analyzed and discussed?
* If a plot is provided, are the axes, title, and datum clearly defined?

The active users are divided into two groups per the ratings they received. The users who received any type of rating (positive, neutral or negative) and the users without any rating feedback from buyers are separated and compared in terms of ‘listing’ and ‘gmv’ features.

## Algorithms and Techniques

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

* Are the algorithms you will use, including any default variables/parameters in the project clearly defined?
* Are the techniques to be used thoroughly discussed and justified?
* Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?

PCA, Feature scaling

As presented in the previous section, the dataset contains 11 features. By using the principal component analysis (PCA), we can reduce the number of features to a reasonable number without losing too much of the integrity of the data. It is an engineer’s discretion to select number of principal components after process the original data using PCA, but it is conventional to select up to three PCs as three dimensional data can be easily visualized and therefore its clusters can also be visually inspected and assessed.

1. Feature selection

Before applying the PCA directly to the original dataset, however, it is recommended to remove or reduce the number of columns by understanding the meaning of each column and the correlations between columns. If any number of columns are strongly correlated, than they can be combined into a single feature. For the dataset in this study, the first and second column, ID and Install Date could be dropped as they only represents random identification numbers of sellers and the usage start date of each seller. Three rating columns (positive, neutral and negative) were initially considered to be irrelevant for analysis because not all active sellers receive ratings and they are given by their buyers and hence they tend to be somewhat inconsistent, but they were added to the final assessment of PCA.

The columns ‘time\_on\_site’ and ‘listing’ were combined into a new column called ‘listing\_per\_day’ to provide how active and how frequent the seller list the materials onto the app.

Now the original 11 features are reduced to 8. (Table 1)

Table 1 List of features after feature selection

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| positive\_rating | newutral\_rating | negative\_rating | listing\_gmv | sale | buyers | gmv | listing\_per\_day |

1. Feature scaling  
   As a final step before the PCA, the data were normalized using sklean MinMaxScaler. This estimator scales and translates each features individually such that the values range from zero to one.
2. Principal Component Analysis

## Benchmark

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

* Has some result or value been provided that acts as a benchmark for measuring performance?
* Is it clear how this result or value was obtained (whether by data or by hypothesis)?

Show silhouette score as a benchmark. Explain what they mean.

If the number of clusters is unknown or labels for clusters are unknown, evaluation after clustering analysis must be performed. The Silhouette Coefficient is an example of the evaluation, where a higher Silhouette Coefficient score corresponds to a model with better defined clusters. The Silhouette Coefficient s for a single sample is then given as:

Where a is the mean distance between a sample and all other points in the same class and b is the mean distance between a sample and all other points in the next nearest cluster. ([http://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient](http://scikit-learn.org/stable/modules/clustering.html" \l "silhouette-coefficient))

# Methodology

## Data Preprocessing

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

* If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?
* Based on the Data Exploration section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?
* If no preprocessing is needed, has it been made clear why?

Remove ‘zero’ data. Combine columns into ‘daily listing’ and etc.

As stated in Section 2, seller records with zero listing has been removed from analysis and they were automatically labeled as ‘New User’ from the app usage point of view. These users could also be split into multiple groups based on the feature ‘time\_on\_site’, but, for the simplicity of the analysis, were assumed to be in one group.

With remaining active user data, I have combined two features of ‘time\_on\_site’ and ‘listing’ to create a new feature ‘listing\_per\_day.’ This combined new feature will be defined by the number of listing divided by the total time after the initial installation of the app for each user. This will reduce the total number of features and normalize the listing numbers with the usage time of the app.

## Implementation

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

* *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
* *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
* *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

Use K-Means and other clustering techniques, then compute silhouette coefficient.  
1. K-means

2. Affinity propagation

3. mean-shift

4. Spectral clustering – may not be proper (because even cluster size)

5. Ward hierarchical

6. Agglomerative clustering

7. DBSCAN

8. Gaussian mixtures

9. Birch

## Refinement

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

* *Has an initial solution been found and clearly reported?*
* *Is the process of improvement clearly documented, such as what techniques were used?*
* *Are intermediate and final solutions clearly reported as the process is improved?*

# Results

## Model Evaluation and Validation

Split data 80/20 (at beginning) and train model with 80% and test model with the rest

## Justification

Using K-means results as ground truth, we could evaluate the goodness of different clustering schemes.

Use F-score?

# Conclusion

## Free-Form Visualization

## Reflection

## Improvement